Validation of reliable 3DTV subjective assessment methodology - Establishing a Ground Truth Database

Jing Li, Marcus Barkowsky, Patrick Le Callet

Validation of reliable 3DTV subjective assessment methodology - Establishing a Ground Truth Database

Jing Li, Marcus Barkowsky, Patrick Le Callet

Subjective assessment methodology for 3DTV

Quality of Experience (QoE) in 3DTV is a multi-dimensional concept which includes image quality, depth quality, and visual comfort. How to measure this multi-dimensional concept is a challenging issue nowadays. In this letter, we introduce a Ground Truth database which is targeted for the standardization of subjective methodologies for QoE in 3DTV.

May the Absolute Category Rating (ACR) be used with 3D stereoscopic content? Experts agree that as long as the degradations are on one single perceptual scale, notably image degradations such as coding artifacts, the previously employed assessment methods such as ACR or DSCQS may be suitable. In 2012, the International Telecommunication Union (ITU) published ITU-R BT.2021[1] for the Subjective assessment methods of stereoscopic 3DTV systems. These recommended methods are derived from ITU-R BT.500 and measure the three primary dimensions of QoE independently: picture quality, depth quality, and visual comfort. However, depending on the transmission conditions, 3DTV may impact all three scales simultaneously. For example, a packet loss may lead to mismatched content in one of the two views leading to an immediate sensation of visual discomfort. In such a case, the previous methods may not be applicable anymore. This is also reflected by ITU-R BT.2021, where the recommended test

[1]International Telecommunication Union - Radiocommunication Sector, "Recommendation ITU-R BT.2021: Subjective methods for the assessment of stereoscopic 3DTV systems", 2012

methods are not suggested for the assessment of naturalness, sense of presence, or the overall QoE. Concerning this issue, in 2013, the IEEE P3333.1[2] Work Group was established to develop novel methods for standardization of subjective quality assessment methodology in 3DTV.

Currently, VQEG experts agreed that the most suitable method for subjective experiments that span several scales is the Paired Comparison (PC) method. Observers just need to choose one sequence in each pair which avoids scale and language interpretation issues; this criterion is also easy to understand. The drawback of PC is the number of visualizations and therefore the length of the subjective experiment, particularly when each pair is visualized in the Full Paired Comparison (FPC) method. To resolve this issue, a new design, Optimized Rectangular Design (ORD),[3] has been proposed to reduce the number of comparisons in PC and is now widely used in the community. In 2014, the ORD was accepted by IEEE P3333.1 Work Group as a standard quality assessment methodology for 3D contents. The basic idea of the ORD method is to arrange the stimuli indices optimally into a rectangular matrix and only compare the pairs within the same row or column. In this way, the number of comparison is significantly reduced compared to FPC.

As it was shown that precision similar to that of FPC can be reached by the ORD method[4], VQEG has therefore decided to run a coordinated subjective experiment on QoE of 3DTV by using the PC ORD method. The obtained results are considered "Ground Truth" for the standardization of subjective assessment methodology for QoE of 3DTV. Thus, the reliability and suitability of ACR, DSCQS or other newly

[2] IEEE P3333.1 WG - Quality Assessment of Three Dimensional Contents based on Psychophysical Studies Working Group, IEEE Computer Society.
[3] J. Li, M. Barkowsky, P. Le Callet, "Subjective assessment methodology for Preference of Experience in 3DTV", IEEE IVMSP, 2013.
[4] J. Li, M. Barkowsky, P. Le Callet, "Boosting Paired Comparison methodology in measuring visual discomfort of 3DTV: performances of three different designs", Proceedings of the SPIE Electronic Imaging, Stereoscopic Displays and Applications, 2013.

designed subjective methods can be evaluated and validated based on this database.

VQEG GroTruQoE3D database

The database is called VQEG GroTruQoE3D (**Gro**und **Tru**th **Q**uality **o**f **E**xperience in **3D**) database. This database contains a well-chosen set of 3D contents (SRC) exhibiting small and large depth budgets, slow and fast planar movement, various kinds of in-depth movement, fine spatial details, strong contrasts, and dark scenes. They were degraded with 18 degradations (HRC) that were selected by experts in order to target a uniform usage of the three scales and their interaction. The distribution of the degradation levels of the selected HRCs on each dimension (image quality, depth quality, and visual comfort) is shown in Figure 1. The video contents are shown in Figure 2. This database has been made available[5] and is currently under evaluation by VQEG's 3DTV group.

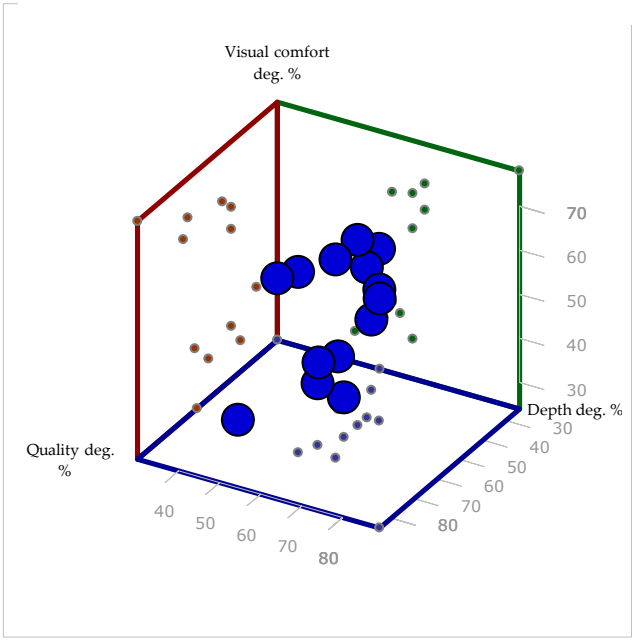

Figure 1. The distribution of degradation levels of the HRCs on each dimension.

[5] ftp://ftp.ivc.polytech.univ-nantes.fr/VQEG_3DTV_GROTRUQOE3D

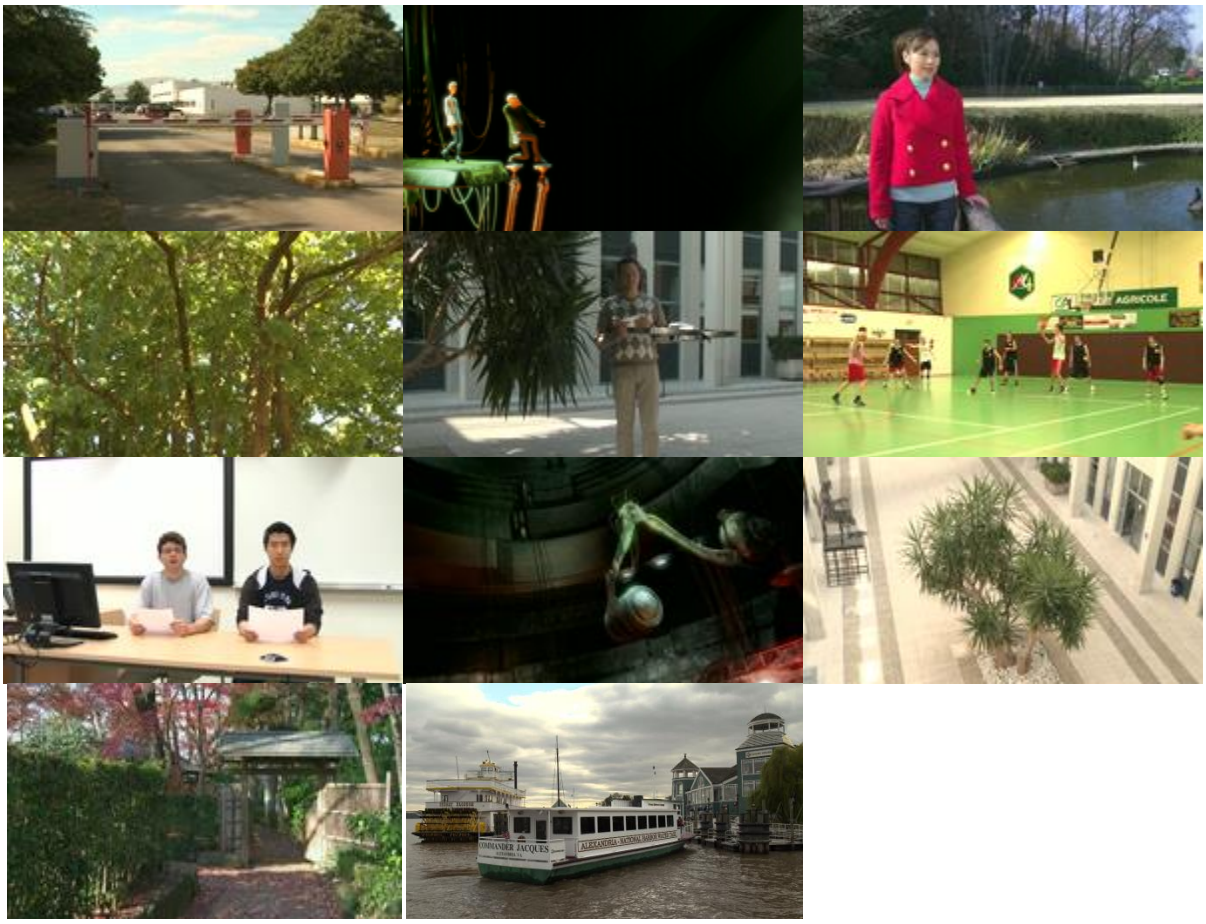

Figure 2. Thumbnails of the VQEG GroTruQoE3D database.

The QoE of the dataset is assessed by the ORD method in such a way that the 18 HRC indices are arranged into a 3 × 6 matrix and only the pairs within the same row or column are compared, which leads to $3 \times \binom{6}{2} + 6 \times \binom{3}{2} = 63$ comparisons per observer. However, considering that there are 11 SRCs altogether, for the ORD method the total number of comparisons would be 11 × 63 = 396 observations per observer, which is still a large number. To make the test feasible, it has been decided to split the workload amongst eight laboratories: IRCCyN (France), INSA (France), Yonsei University (Korea), UPM (Spain), NTIA (USA), T-labs (Germany), FuB (Italy) and BskyB (UK). The construction of a common set of pairs for all

labs is required which will allow for the validation of results among labs.

The common set includes two fixed SRCs and 18 HRC pairs. The two SRCs are selected in such a way that they are sensitive to test environment. The 18 HRC pairs are constructed by a 3 × 3 HRC matrix which is a subset of the whole 3 × 6 HRC matrix. The selected 9 HRCs represent 3 levels in 3 dimensions of the 3D QoE, i.e., image quality, depth quality, and visual comfort.

The obtained data will be collected and then analyzed for two main purposes. The first goal is to evaluate the validity of the acquired data in the different subjective assessment labs, thus allowing the creation of a large common dataset. When two alternative forced choice (2AFC) Paired Comparison is used as the assessment methodology, scale adaptation problems do not arise. The second goal is to establish a scale value for each Processed Video Sequence (PVS). This eases the comparison of the results to assessment methods that use direct scales such as Absolute Category Rating or Double Stimulus Continuous Quality Scale. For the first goal, the main statistical analysis tool used is Barnard's-exact-test, which examines whether the PC preference data obtained from two labs is significantly different. Thus, "outliers" may be detected by determining a threshold on the total number of significantly different pairs. For the second goal, to convert the paired comparison data to scale values, the Bradley-Terry model will be applied. This could provide not only the scales for all PVSs, but also some statistics, including the confidence intervals for each PVS, how well the model fits, etc.

Validation of reliable subjective assessment methodology in 3DTV

The results of the GroTruQoE3D evaluation may be used to verify not only the performance of existing subjective quality

Jing Li *received her M.S. degree in Electronic Engineering from Xidian University, Shaanxi, China, in 2010, and her Ph.D. degree from University of Nantes, France, in 2013. Her research interests include subjective assessment methodologies and objective modeling on QoE, visual discomfort and quality of experience in 3DTV and Ultra HDTV. She is now leading the construction of the VQEG GroTruQoE3D database.*

Marcus Barkowsky *received the Dr.-Ing. degree from the University of Erlangen-Nuremberg in 2009. He joined the Image and Video Communications Group at IRCCyN at the University of Nantes in 2008, and was promoted to associate professor in 2010. His activities range from modeling effects of the human visual system, in particular the influence of coding, transmission, and display artifacts in 2D and 3D to measuring and quantifying visual discomfort and visual fatigue on 3D displays using psychometric and medical measurements. He currently co-chairs the VQEG "3DTV" and "Joint Effort Group Hybrid" activities.*

assessment methods, but also the impact of different perceptual measurement scales, the influence of observer training on the results, etc. New methodologies may be developed based on the results.

The existing quality assessment methods can be validated, for example, using the following criteria:

1) Correlation analysis: By calculating the Pearson Linear Correlation Coefficients (PLCC) and Spearman Rank Order Correlation Coefficients (SROCC), the correlation between the results obtained by Pair Comparison (Ground Truth) and the tested methodology can be obtained, which shows the consistency of the tested methodology with the ground truth.

2) Accuracy analysis: By calculating the Root Mean Square Error (RMSE) between the ground truth and the fitted data, the accuracy of the tested methodology can be evaluated.

3) Distinguishability analysis: The distinguishability of Pair Comparison can be tested by the Barnard's-exact-test, where the significance of the observer's preference on each pair can be shown. For the tested methodology, the distinguishability can be evaluated by confidence intervals or Student's t-test. Another possible way is to convert the results of the tested methodology to PC data and then Barnard's-exact-test may be used. Statistical analysis between the two subjective test methodologies is enabled and the relative performance of the tested methodology can be evaluated.

With this GroTruQoE3D database, a list of verified and validated assessment methods for 3DTV may be established for standardization in ITU Recommendations.

Patrick Le Callet *is a full Professor at Ecole polytechnique de l'Universite de Nantes. Since 2006, he is the head of the Image and Video Communication lab at CNRS IRCCyN, a group of more than 35 researchers. He is mostly engaged in research dealing with the application of human vision modeling in image and video processing. He is currently co-chairs the "3DTV" activities and the "Joint-Effort Group", driving mostly High Dynamic Range topic in this latest. He is currently serving as associate editor for IEEE transactions on Circuit System and Video Technology, SPIE Journal of Electronic Imaging and SPRINGER EURASIP Journal on Image and video Processing.*